2015-03

IP infrastructure geolocation

Cai, Guan Yan

Monterey, California: Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

IP INFRASTRUCTURE GEOLOCATION

by

Guan Yan Cai

March 2015

Thesis Advisor:	Robert Beverly
Second Reader:	Geoffrey Xie

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	03-27-2015	Master's Thesis 09-16-2013 to 03-27-2015

4. TITLE AND SUBTITLE

IP INFRASTRUCTURE GEOLOCATION

5. FUNDING NUMBERS

N66001-2250-58231

6. AUTHOR(S)

Guan Yan Cai

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Department of Homeland Security
245 Murray Lane SW, Washington, DC 20528

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The views expressed in this document are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol Number: N/A.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(maximum 200 words)*

Physical network maps are important to critical infrastructure defense and planning. Current state-of-the-art network infrastructure geolocation relies on Domain Name System (DNS) inferences. However, not only is using the DNS relatively inaccurate for infrastructure geolocation, many router interfaces lack DNS name entries. We adapt the technique of Wang *et al.* to send traceroute probes from distributed vantage points, and approximate a target's location by finding the nearest landmark. To evaluate the technique's performance, we geolocate router interfaces previously geolocated via DNS-based router positioning (DRoP). Our results show that 50% of the targets have error distances greater than 2,400 km; however, 75% of the nearest landmark predictions are less than 5 ms distant. We find that geolocation accuracy is insensitive to vantage point location, while the use of more vantage points improves accuracy. To better understand these results, we use Constraint-based Geolocation (CBG) on a subset of DRoP predictions. Forty-six percent of 4,638 DRoP location inferences are in regions outside the feasible physical boundaries imposed by CBG and 56% are 1,800 km away from the CBG centroid. Our findings suggest that our methodology can supplement prior work to not only geolocate infrastructure without DNS names, but also improve accuracy.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Internet, IP geolocation, IP infrastructure, routers, Domain Name System	77
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

IP INFRASTRUCTURE GEOLOCATION

Guan Yan Cai
Civilian, Ministry of Defense, Singapore
B.S., National University of Singapore, 2006

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
March 2015

Author: Guan Yan Cai

Approved by: Robert Beverly
Thesis Advisor

Geoffrey Xie
Second Reader

Peter J. Denning
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Physical network maps are important to critical infrastructure defense and planning. Current state-of-the-art network infrastructure geolocation relies on Domain Name System (DNS) inferences. However, not only is using the DNS relatively inaccurate for infrastructure geolocation, many router interfaces lack DNS name entries. We adapt the technique of Wang *et al.* to send traceroute probes from distributed vantage points, and approximate a target's location by finding the nearest landmark. To evaluate the technique's performance, we geolocate router interfaces previously geolocated via DNS-based router positioning (DRoP). Our results show that 50% of the targets have error distances greater than 2,400 km; however, 75% of the nearest landmark predictions are less than 5 ms distant. We find that geolocation accuracy is insensitive to vantage point location, while the use of more vantage points improves accuracy. To better understand these results, we use Constraint-based Geolocation (CBG) on a subset of DRoP predictions. Forty-six percent of 4,638 DRoP location inferences are in regions outside the feasible physical boundaries imposed by CBG and 56% are 1,800 km away from the CBG centroid. Our findings suggest that our methodology can supplement prior work to not only geolocate infrastructure without DNS names, but also improve accuracy.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AP Access Point

API Application Programming Interface

AS Autonomous System

Ark Archipelago

BSSID Basic Service Set Identification

CAIDA Cooperative Association for Internet Data Analysis

CBG Constraint-Based Geolocation

CDF Cumulative Distribution Function

CLLI Common Language Location Identifier

DNS Domain Name System

DRoP DNS-based router positioning

FQDN Fully Qualified Domain Name

GPS Global Positioning System

IATA International Air Transport Association

ICAO International Civil Aviation Organization

ICMP Internet Control Message Protocol

IETF Internet Engineering Task Force

IP Internet Protocol

IPv4 Internet Protocol Version 4

IPv6 Internet Protocol Version 6

ISP Internet Service Provider

ITDK Internet Topology Data Kit

LIS Location Information Server

NPS Naval Postgraduate School

PC Personal Computer

POP Point of Presence

PTR Pointer Record

RIR Regional Internet Registry

RTT Round Trip Time

SSID Service Set Identification

TCP Transmission Control Protocol

TTL Time To Live

UDP User Datagram Protocol

U.S. United States

VoIP Voice Over Internet Protocol

WPS WiFi-Based Positioning System

Acknowledgments

I would like to thank Professor Robert Beverly for his expert guidance on IP geolocation, as well as his feedback on writing the thesis. He was always patient when I shared with him the problems that I had encountered in my research. His directions and encouragement gave me strength to persevere in the right direction. It has been my pleasure to have worked with Professor Beverly.

I would like to express my gratitude to Professor Geoffrey Xie for his insightful feedback. He identified confusing areas and suggested ways to make my thesis easier to understand. I am grateful to Professor Justin Rohrer for his ideas, especially on handling negative estimated distances, and developing ArkQueue, which allowed my trace route measurements to complete faster. Mr. Michael McCarrin provided deep insights on applying CBG to evaluate the accuracy of ground truth. His questions identified flaws in my approaches and that ensured my work was on the right track.

Finally, I would like to thank my wife for taking care of the family while I stayed up late in school to write my thesis. I would also like to thank my son for understanding that I could not play with him because I had to work on my thesis.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

Geolocation is the determination of the position of an object in physical space. The result of geolocation is generally a meaningful location such as a street address. In contrast, positioning produces a more precise geographic coordinate.

When the Global Positioning System (GPS) is unavailable, objects can fall back on geolocation services to deduce or infer their current geographic location. One of the popular techniques to geolocate is to analyze received radio waves, which is known as radiolocation.

Since network-enabled devices have network addresses that are indicative of their attachment point, these addresses can be used for geolocation. For example, Internet devices can use a device's globally unique Internet Protocol (IP) address to estimate its location. Various IP geolocation techniques are discussed in Chapter 2.

Internet devices can be broadly classified as either "edge" or "infrastructure." Edge devices are the end points on the Internet where traffic originates or terminates. Examples of edge devices include Personal Computer (PC), or smart phones. Infrastructure devices support Internet communication between edge devices by relaying information from one edge device to another. Examples of infrastructure devices include switches, routers, and firewalls.

Differences in capabilities between edge and infrastructure devices means geolocation techniques designed for edge devices may be unsuitable for infrastructure devices. Edge devices generally have hardware components that can provide geolocation information while infrastructure devices do not. For example, GPS or WiFi-Based Positioning System (WPS) is unsuitable for network routers because they are not equipped with wireless receivers and cannot easily receive signals within a datacenter. Unlike edge devices, infrastructure devices are usually immobile and remain in the same area. The geolocation of routers and access points generally do not change on short time scales.

Further, commercial IP geolocation efforts have largely ignored infrastructure geolocation. Instead, most geolocation research and commercial services focus on the edge. For example, geolocation of edge devices permits targeted location-based advertising. Similarly, geolocation can enforce geographic access control, for instance to determine if a user has rights to a company's digital contents. The National Broadcasting Company provided free live streaming of the XLIX Super Bowl for viewers within the United States (U.S.) and Mexico, but required viewers to purchase a subscription if they are outside the two countries [1].

This thesis introduces and examines a new technique that deduces a router interface geographic location using topology information, delay measurements, reference points, which are router interfaces with known locations. Infrastructure devices such as router interfaces generally have fixed locations, static network addresses, consistent network topology information. The characteristics of router interfaces and the estimated proximity to reference router interfaces can be leveraged to estimate their location. Close proximity of router interfaces to reference points may support highly-accurate, fine-grain geolocation of router interfaces.

In the next section, we discuss the importance of designing and developing techniques that accurately geolocate infrastructure devices.

1.1 Motivation

Geolocation of infrastructure devices allows better planning that improves the resilience of critical infrastructure. Many of today's critical services operate on the Internet or IP infrastructure, such as electric distribution grids, medical, financial, transportation and communication services. To be resilient against natural disasters or terrorist attacks, IP infrastructure should be planned with redundancy in mind. Achieving redundancy in the IP infrastructure of a given area allows the services in that area to continue operation in times of crisis without crumbling under load. Therefore, to determine if the IP infrastructure is robust in a particular area, planners require the geographic location of Internet infrastructure devices in that area.

In addition to infrastructure planning, router geolocation helps to reveal Point of Presence (POP) and facilitates the generation of POP-level maps of the Internet. POP-level maps

show the size of each Autonomous System (AS) network in terms of the number of routers, their connectivity, and the number of colocations of the network. Details from POP-level maps help identify critical nodes of the network as well as to understand the dynamics of a network at a given area [2].

1.2 Research Questions

This thesis examines the accuracy of our technique of using topology information, delay measurements, and reference points to deduce the geographic location of router interfaces. In this study, we investigate the following:

1. What is the accuracy of geolocating router interfaces using the our technique?
2. What are the factors that affect the accuracy of our technique?
3. Is the ground truth accurate?

1.3 Significant Findings

1. To be best of our knowledge, this is the first work to geolocate router interfaces that do not have hostname.
2. Our methodology determines landmark router interfaces that are near the target router interfaces. Seventy-five percent of targets have estimated Round Trip Time (RTT) of less than 10 ms away from their nearest landmarks.
3. We evaluated the accuracy of DNS-based router positioning (DRoP) across its locations and found that only 54% of the router interfaces assigned with DRoP locations are also within the CBG possible region.
4. We discovered examples of incorrect DRoP geolocations and determined the errors were due to bad DNS inferences.
5. We found that geolocation accuracy was insensitive to the location of the vantage point but more vantage points improves accuracy.

1.4 Thesis Structure

The remainder of the thesis is organized as follows:

- Chapter 2 surveys available IP geolocation techniques and related work.

- Chapter 3 discusses the methodology of geolocating IP addresses by using router interfaces as landmarks. In this Chapter, we also outline the procedures of evaluating the accuracy ground truth with the Constraint-Based Geolocation (CBG) technique.
- Chapter 4 examines and analyzes the results of geolocating router interfaces with our methodology. We analyze the accuracy of using our methodology to geolocate global router interfaces, and in North America and Oceania continents. We also explore the impact on geolocation accuracy when we vary the location of the vantage points, as well as the number of vantage points involved in the process. To understand the causes of inaccurate results, we highlight three examples of our geolocation results. As the accuracy of our methodology depends on the accuracy ground truth, we analyze the results of geolocating some of our ground truth data with CBG.
- Chapter 5 concludes our work and recommends potential areas of research in IP infrastructure geolocation.

CHAPTER 2:
Background and Related Work

Obtaining geographic information of IP addresses from network operators is unsuitable. Network operators are allocated with IP address ranges by their respective Regional Internet Registrys (RIRs). They then assign these IP addresses to their network resources or allocate to their customers. While network operators have the geographic information of their network resources or customers, they generally do not answer geolocation queries from the public. They may be restricted by company policies to ensure customer security and privacy are not compromised, and to reduce administrative overheads. On the other hand, requesting geolocation information of multiple IP addresses belonging to different network operators is slow and inefficient. To request for geographic information of an IP address, the public has to determine the relevant network operator then submit the request. This method does not scale up for large number of IP geolocation requests. Clearly, more efficient IP geolocation solutions are required.

Geolocating devices on the Internet is non-trivial. The Internet is a collection of many independent, inter-connected networks with physical presence around the world. The most fundamental piece of information about an Internet host, its IP header, which reveals no information about its geographic location. Clearly, the Internet was not designed with geolocation in mind.

The Internet Engineering Task Force (IETF) has realized many applications require location information and doing so may have security and privacy implications, so they established the GEOPRIV working group to develop and refine location representations in IETF protocols without compromising on privacy and security [3]. Examples of enhancements include RFC3825 [4] and RFC4776 [5], and defining protocols to discover the local Location Information Server (LIS) [6].

Before IETF protocols are enhanced with geolocation features, the location of an Internet device has to be deduced by other means. Internet devices that want to geolocate themselves may collect and analyze information about its environment. To geolocate a target

Internet device other than itself, it can query IP geolocation databases with target device's IP address (Section 2.1), or collect delay measurements between that device and a reference point (Section 2.2), or consider topology information of that device (Section 2.3).

Edge devices in WiFi-enabled environments may rely on WPS for geolocation. WPS devices geolocate their location by querying public WiFi location databases with the Basic Service Set Identifications (BSSIDs) and Service Set Identifications (SSIDs) of nearby WiFi access points, then together with measured received signal strength of the Access Points (APs) and results of queries from the location databases, the device is able to calculate its location. WPS is considered a *target-assisted* geolocation technique because the geolocation process is voluntarily initiated and facilitated by the edge device itself.

Domain Name System (DNS) Pointer Records (PTRs) of IP addresses may hint the geographic location of the Internet device; however, the PTRs of IP addresses may not indicate geographic information and the assignment of PTRs to IP addresses is not mandatory. Zhang *et al.* found from their experiment that 0.5% of router interface IP addresses were assigned with incorrect DNS PTRs [7]. This caused the hostnames returned from PTR queries to be incorrect, hence the router interfaces were *misnamed*. Zhang *et al.* attributed router interface misnaming to be caused by administrative mistakes or the lack of timely updates. Zhang *et al.* used trace routes to discover misnamed router interfaces. In addition to misnamed router interfaces, router interfaces may not be assigned with PTRs. This prevents the router interface hostnames from being identified. Of the 31,790K router interfaces surveyed by Huffaker *et al.*, they found 18,956K (40.4%) routers without hostnames [8]. The inability to query accurate router interface hostname or makes geolocation based on hostname inference difficult.

A *whois* of the IP address reveals administrative information including the geographic address of the IP address, but the device assigned with that IP address may not reside at that location. The whois protocol [9] allows a user to query a whois server to obtain information of an Internet resource, which can be an IP address block, domain name, or an autonomous system. The information given in whois may not indicate the true physical location of that IP address. For example, the whois information for 205.155.65.20 (www.nps.edu) does not indicate its true location. The information states that the IP address belongs to California State University, and the given geographic address is 401 Golden Shore, Long Beach, CA.

In general, *non-target-assisted* IP geolocation techniques may be split into two phases. Non-target-assisted geolocation refers to geolocation of a target Internet devices without them directly contributing geographic-related information about their location. The first phase is to compile a list of geolocation information to be associated with IP addresses. The geolocation information can be fine-grained, such as geographic coordinates or street addresses; or coarse-grained, such as cities or states. One of the popular sources of free geolocation information is GeoNames [10]. The second phase is to associate IP addresses to their correct geolocations using various geolocation techniques.

IP Geolocation techniques can be broadly categorized according to the data used to drive geolocation. In this thesis, we group them into three categories: data-based, delay-based or topology-based.

This chapter surveys the popular geolocation techniques of each category and discusses the limitations of each technique.

2.1 Data-Based Geolocation

Geolocation driven by databases is the earliest form of IP geolocation [11]. Information from whois [12] and DNS [13] are used to create rudimentary IP geolocation databases, which supports coarse-grained IP geolocation.

Data-based geolocation returns the geographic location of an IP address or range by referring to the previously determined geographic location i.e., the geographic data of that IP address or range from a geolocation database. The geographic data for that IP address or range may consist of one or more of the following: country, state, city, ZIP code, latitude and longitude. Geolocation databases are provided by either commercial or academic entities.

The strength of data-based geolocation is quick response and high efficiency. As measurements and processing have been precomputed, geolocation information is a simple task of looking it up from the database.

Companies provide different levels of geolocation services, with some services being free of charge and others requiring expensive subscriptions [14]. Free IP geolocation databases

are offered by HostIP [15] and IPInfoDB [16] while MaxMind [17] and IP2Location [18] charge several of hundreds of U.S. dollars for access to their databases.

HostIP [15] and Spotter [19] are geolocation databases provided by non-commercial entities. HostIP is a community-driven geolocation service. It provides an Application Programming Interface (API) for participating users and ISPs to provide direct feedback regarding their geolocation. HostIP does not specialize in infrastructure geolocation. Spotter is a research project by the Eotvos Lorand University in Budapest, Hungary. Spotter uses both delay-based and topology-based geolocation techniques to populate its database of geolocation information. Laki *et al.* measured the accuracy of Spotter with 23,000 network routers from Cogent. They found that Spotter could geolocate 35% of the routers within 10 km, and almost 70% of the routers have error distances below 50 km [19].

The methodology used by most geolocation database service providers to populate their databases is unknown. Unlike non-commercial geolocation databases such as HostIP and Spotter, commercial geolocation services do not disclose their underlying techniques used to collect geolocation information of IP addresses. Within each company, their geolocation services are further differentiated into tiers, each offering different levels of accuracy and recency. The methodologies used between each tier are also not disclosed by the companies.

Shavitt and Zilberman compared the accuracy [20] of GeoBytes [21], HostIP, IP2Location, IPligence [22], MaxMind, NetAcuity [23], and Spotter [19], with ground truth from Cooperative Association for Internet Data Analysis (CAIDA) [24]. The ground truth dataset consisted of 25K IP addresses from a tier-1 Internet Service Provider (ISP), a tier-2 ISP, and five research networks. The accuracy of each database can be determined from the "City Match" column. For a geolocated IP address to be considered a city match, it has to have an error distance between itself and the ground truth of less than 100 km. The free geolocation database from HostIP did not perform well. Only 28.1% of the IP addresses in HostIP database are also found in the ground truth. Of the 28.1% can be compared with the ground truth, HostIP located only 17.9% in the correct city. Commercial geolocation databases IP2Location and IPligence did not perform better. While 93.9% of their IP addresses can be found in the ground truth, less than 15% matched the correct city. Their results are summarized in Table 2.1.

Database	Matched IP Addresses	Country Match	City Match
GeoBytes	67.3%	80.1%	26.5%
HostsIP	28.1%	89.0%	17.9%
IP2Location	93.9%	80.9%	14.16%
IPligence	93.9%	81.0%	0.8%
MaxMind	79.6%	84.7%	29.4%
NetAcuity	67.9%	96.9%	79.1%
Spotter	54.1%	85.6%	27.8%

Table 2.1: Results of Shavitt and Zilberman comparison [20] with CAIDA's ground truth dataset [24].

2.2 Delay-Based Geolocation

Delay-based techniques try to deduce a target's geolocation primarily by actively or passively measuring the end-to-end delays between the target and other Internet resources with known locations. The host on the Internet that initiates a delay measurement is called a *vantage point* while hosts on the Internet that have known geographic location information and are used as reference points are called *landmarks* (Figure 2.1).

The end-to-end delay, $D_{end-to-end}$, between two points on a packet-switched network like the Internet can be characterized by Equation 2.1.

$$D_{end-to-end} = d_{trans} + d_{prop} + d_{proc} + d_{queue} \tag{2.1}$$

Where,

- d_{trans} is the transmission delay. It represents the amount of time required to send a complete packet content onto the transmission medium.
- d_{prop} is the propagation delay. It represents the amount of time required for a bit of a packet to travel between two points on the network.
- d_{proc} is the processing delay. It represents the amount of time required by the router to process the header of the packet.
- d_{queue} is the queuing delay. It represents the amount of time the packet waits in the queue of the output interface, when it waits for turn to be transmitted.

Delay-based geolocation that considers end-to-end delay is usually dominated by the propagation delay, d_{prop}, and the transmission delay, d_{trans}. Queuing delay, d_{queue}, can be minimized by sending a series of delay measurement probes. For simplicity, we assume the processing delay, d_{proc}, and queuing delay, d_{queue}, are negligible. Typically, end-to-end delay measurements use using small packets. This means the transmission delay is small. For example, a 1,500-byte packet requires 0.12 milliseconds to be transmitted over a 100 Mbps link. When two points are geographically near each other, then the propagation delay becomes small. This is common for fine-grained, delay-based geolocation. However, when the two points are geographically distant points, then the propagation delay becomes significant. For example, if a vantage point and the target are on opposite coasts of U.S., separated by a distance of approximately 5,000 km, and the packet travels through optical fiber that has a wave propagation speed of two-thirds the speed of light, then the propagation delay is 25 milliseconds.

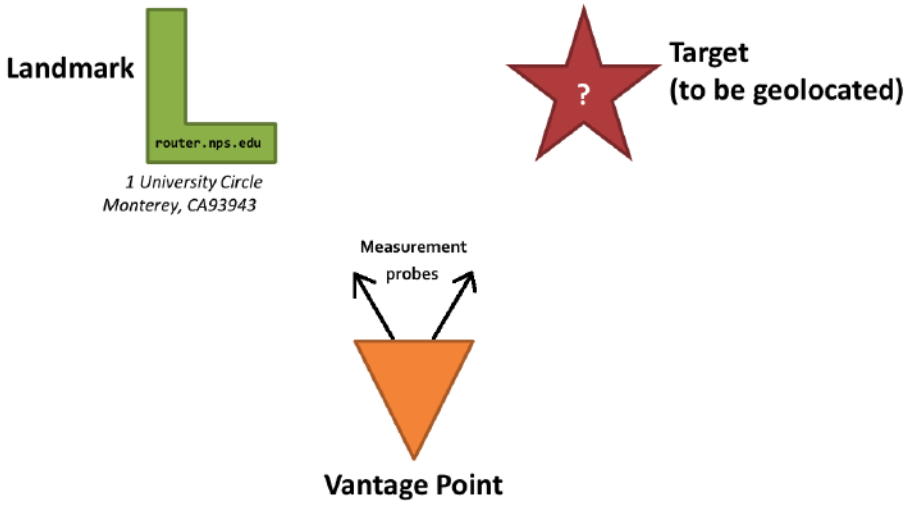

Figure 2.1: A vantage point sends out measurement probes to a landmark, router.nps.edu at 1 University Circle, Monterey, CA 93943, and target with unknown location.

The advantage of delay-based geolocation is that the estimated target location is bounded by physical constraints. By assuming that the probes travel at two-thirds the speed of light, the measured delay limits the furthest possible geographic distance that the target can be away from the vantage point. The ability to constrain location estimates differentiates delay-based geolocation from data-based geolocation. The error estimates of data-based

geolocation is unbounded due to inaccurate location information derived from whois or DNS databases.

GeoPing [13] is one of the first delay-based geolocation technique to make use of the relationship between end-to-end delay and geographic distance to geolocate an Internet host. Padmanabhan and Subramanian found the correlation between end-to-end delay and geographic separation. They leveraged this relationship to estimate the geographic location of a target. Their technique constructs a delay map with each entry of the map holding the location of a landmark, and a delay vector that records the delay from this landmark to each vantage points. A delay vector is created for the target that records the delay from the target to each vantage point. The technique determines the nearest landmark by finding the least Euclidean distance between the target's delay vector and each entry in the delay map. The location of the target is estimated as the location of its nearest landmark. Padmanabhan and Subramanian noted that as the number of landmarks increases, the likelihood of finding a nearby landmark increases, which leads to better geolocation accuracy.

CBG [25] uses multilateration to infer the geographic location of Internet hosts. Multilateration refers to the use of multiple distance measurements between vantage points and the target to deduce the location of the target. Prior to target geolocation, the technique determines the *best line* for each vantage point by measuring the delay from each vantage point to all other vantage points. The best line describes the relationship between delay and distance for the particular vantage point. CBG then calculates the delay between each vantage point and target. The delay is mapped to the respective best line to convert delay into distance estimates. This distance estimate is the upper bound of the furthest possible distance between the target and the vantage point. It can be thought of as the radius of a circular area representing the possible location of the target. Each area is then intersected with one another to further constrain and reduce the possible region of the target. Unlike GeoPing, which geolocates a target by producing a set of discrete, possible locations, CBG geolocates the target by producing a contiguous area of the possible location of the target.

Finally, *Posit* [26] is a delay-based geolocation technique that requires a relatively small number of probe packets and estimates the location of a target with the "statistical embedding" technique. The technique first constructs two delay vectors: one for the delay between the set of vantage points and a set of targets, and one for the delay between the

set of vantage points and a set of landmarks. Then a training set of targets with known locations is used to derive the probability distribution of distances between the target and the set of landmarks. Finally, each target from the set of targets is geolocated using the statistical embedding algorithm. The algorithm uses CBG to derive a possible region of the target location, then applies the constrained region to the trained likelihood distribution to obtain the most likely location of the target.

2.3 Topology-Based Geolocation

Topology-based geolocation techniques rely on topology information to estimate the geographic location of a target host. These techniques assume that addresses of Internet hosts or routers that are topologically close are also geographically close. While these techniques geolocate targets primarily with topology information, topology information is usually obtained from databases or delay-based measurements. The hostname of the target may be queried from the DNS database to reveal the topology of the target, and may also encode geographic information about the target. Router interfaces may be named after airport codes, city names, or creative abbreviations of city names. For example, the hostname of the router interface `ccr21.par01.atlas.cogentco.com` contains three hints of its possible geographical location [8]. The strings that "ccr" may indicate the airport code for Concord, CA; "par" may indicate Paris, France, or one of the 20 towns named Paris in U.S.; and "atlas" may indicate the city Salas Atlas in Spain. Trace routes to the target show the path that network packets travel from each vantage point to the target. The delay measured at each "hop" of the path provides complementary information about the proximity between two adjacent hops.

TBG or topology-based geolocation, optimizes the CBG technique by obtaining and considering topology information relating to a target in addition to performing multilateration to constrain the possible area of the target [27]. Basset *et al.* highlights that multilateration techniques such as CBG, perform well when at least one of the vantage points is near the target. This characteristic of multilateration techniques implies inconsistent geolocation accuracy because in situations with vantage points far from the target, geolocation becomes inaccurate. TBG obtains the topology of the vantage points by performing trace routes amongst the landmarks. The topology information obtained is then used to minimize the error when geolocating the target.

Web servers can be used as landmarks to achieve highly accurate and reliable location estimates of targets [28]. Wang *et al.* conceived of a geolocation methodology that is able to geolocate targets down to the street-level i.e., fine-grain geolocation. Wang *et al.* calls this technique *street-level, client-independent* geolocation. The method first finds web servers that publish web pages containing their geographic locations and use these web server as landmarks. Then, it performs trace routes from vantage points to these landmarks and the geolocation target to discover the topology of the target. Finally, the target is assumed to have the same geolocation as its nearest landmark. The limitation of this technique is that it requires web servers to reside in the same compound as the geographic address published on their web pages. Since companies may prefer to co-locate web servers in data centers, or host their web pages on the cloud, this method may not find sufficient web-based landmarks that reside in the same location as their published geographical addresses.

The geolocation techniques that we discussed above were not tailored to geolocate infrastructure devices. As these techniques do not exploit the characteristics of routers for more accurate geolocation, they may not be ideal to perform fine-grained geolocation of infrastructure devices. We adapted the street-level, client-independent methodology by Wang *et al.* to perform fine-grained infrastructure geolocation by using router interfaces instead of web servers as landmarks. The intuition behind our methodology is that many infrastructure devices such as routers, typically colocate in a datacenter; if one or more routers with known location (landmarks) colocate in the same datacenter as the router with unknown location (target), then it is possible for our methodology to obtain a highly-accurate location estimate of the target. We discuss our methodology in detail in Chapter 3.

DRoP or DNS-based router positioning, is a topology-based geolocation technique that specifically caters to infrastructure geolocation [8]. Huffaker *et al.* notes that certain router interfaces have well-structured Fully Qualified Domain Names (FQDNs) that may contain hints to their geographic location, such as the city or airport names. To make use of hints from FQDNs of router interfaces for geolocation, Huffaker *et al.* creates the location hint dictionary, which is composed from International Air Transport Association (IATA) and International Civil Aviation Organization (ICAO) codes, Common Language Location Identifier (CLLI) code database, United Nations Code for Trade and Transport Locations, and major city names. Each substring of the FQDNs that matches the location hint is called

a geographic hint. Huffaker *et al.* then builds router sets by selecting routers from the July Internet Topology Data Kit (ITDK) 2013 dataset that have at least one substring matching the location hint dictionary. A *geohint* contains the public suffix and geographic hint of a router hostname from the router set, and the geographic coordinates. Each router set is associated to a geohint by the matching geographic hint. To ensure that the geohints are accurate, Huffaker *et al.* probes each router interfaces of each router set to calculate the standard deviations of the minimum RTT and hop count of the set. They used ground truth from the operators of six public suffixes as training set to create the classifier of geohint validity. The standard deviations of the minimum RTT and hop count of each router set are used to determine the validity of each geohint. Finally, with the geohints database ready, it generates general geolocation rules that map substrings of hostnames to a geohint. The association between a router hostname and a geohint enables the geographic location (country, city and geographic coordinates) of a router interface to be inferred from its FDQN.

There are two limitations with DRoP. First, it is unable to geolocate router interfaces that are not assigned with FQDNs. Of the 31,790K router interfaces analyzed by DRoP, 40.4% of the router interfaces do not have hostnames. Second, DRoP is unable to geolocate router interfaces with FQDNs that do not match with the geohint database. Of the 59.6% of the router interfaces that have FQDNs assigned, 57.5% of them do not matched the geohint database. In total, DRoP is unable to geolocate 45% of the router interfaces. This thesis aims to address the limitation of DRoP. We discuss our methodology that is able to geolocate router interfaces without relying on their hostnames in Chapter 3.

CHAPTER 3:
Methodology

The methodology of our fine-grained router interface geolocation is an adaptation of the street-level geolocation technique [28] that aims to overcome the limitations of DRoP [8] in router interface geolocation. Whereas the prior street-level geolocation work used web servers that publish their geographic addresses as landmarks, we rely on router interfaces with known geographic locations. The intuition behind leveraging router interfaces with known location as landmarks to geolocate other router interfaces is simple: routers are often co-located, for instance in large data centers and exchange points. We therefore hypothesize that it is reasonable to expect to find landmarks with low latencies to the router interface targets.

Our methodology has three main steps:

1. First, we construct the list of landmarks and targets with the updated router interface geolocation datasets from DRoP (Section 3.2).
2. Next, we use traceroutes from Ark vantage points to actively measure the forward path and forward path round-trip delays to each router interface in the landmark set and each in the target set (Section 3.3).
3. Finally, we generate the estimated delay vectors for each target (Section 3.4). The estimated geolocation of the target router interface is the location of the landmark from the estimated delay vector with the minimum estimated delay.

In this Chapter, we first introduce the Archipelago (Ark) active measurement platform that is used to perform traceroutes to landmarks and targets (Section 3.1), then we detail our three-step fine-grained router geolocation methodology. We highlight the limitations of our methodology (Section 3.5), and finally, we explain the procedures of using CBG to determine the accuracy of the router interface geolocation datasets (Section 3.6).

3.1 The Ark Active Measurement Platform

The Ark infrastructure is CAIDA's distributed active measurement platform to conduct experiments for Internet research. We use the Topology-on-Demand (ToD) [29] service on Ark to send probes to perform traceroute that identify the path between an Ark monitor and an Internet host, as well as the RTTs of the intermediate hops and the destination. Traceroutes and RTT measurements are generally useful for research on Internet traffic, topology, routing, and performance [30].

Each traceroute request that is submitted to the Ark measurement infrastructure specifies a source and a destination. The source of the traceroute is an Ark monitor i.e., a vantage point. In our work, the traceroute destination is either a router interface with known location i.e., landmark, or router interface with unknown location i.e., target.

3.1.1 Selecting Vantage Points

Vantage points are selected according to their geographic location and availability. The Ark infrastructure supports a total of 82 monitors distributed across the world that we can select as vantage points for our traceroute measurements. In our methodology, one vantage point from each continent is selected for the experiment. Selecting vantage points from each continent minimizes the risks of having traceroutes that fail to reach landmarks or targets. If probes from one continent are dropped due to unexpected congestions then sending probes from other continents may allow probes to take a different route that bypasses the congestion and successfully reach the intended landmark or target. Vantage points have to be available and operational during the entire traceroute collection process. If a vantage point goes offline while our measurements are on going, then we have to replace results from this vantage point by either repeating the measurements from the same vantage point when it comes online, or by performing the measurements from a more reliable vantage point. Our implementation expects the selected vantage points to be operational as it does not repeat measurements on a vantage point that returns online, or repeat measurements from another vantage point. Therefore, the selected vantage points have to be operational and reliable.

We perform five trial traceroute measurements to 8.8.8.8 from the available Ark monitors, then for each continent, we select the vantage point that produces the lowest RTT to 8.8.8.8.

The list of selected vantage points are shown in Table 3.1. The sensitivity of our vantage point selection to geolocation accuracy is examined in Section 4.3.

Name	Location	Continent
san-us	San Diego, CA, U.S.	North America
sin-sg	Singapore	Asia
mel-au	Melbourne, Australia	Oceania
cbg-uk	Cambridge, United Kingdom	Europe
sao-br	Sao Paolo, Brazil	South America
pry-za	Pretoria, South Africa	Africa

Table 3.1: Vantage points that are selected for the experiment.

3.2 Building the Landmarks and Targets Lists

The landmarks and targets lists are extracted from the October 2014 DRoP geolocation results. The results consist of two datasets: a set of 6,041,769 router interface IP addresses from CAIDA's ITDK nodes dataset that are geolocated by DRoP [31], and a set of 8,141 possible locations of these router interfaces [32].

Each row of the router interface IP address dataset is defined as follows:

$$< LocationID, DataSource, IPaddress_i, IPaddress_{i+1}, ..., IPaddress_{i+n} >$$

For example, router interfaces 75.102.5.183, 75.102.17.42, and 75.102.5.186 have the *LocationID* 73, where the location is inferred via CAIDA's DRoP algorithm.

Each row of the router interface location dataset is defined as follows:

$$< LocationID, Country, Region, City, Latitude, Longitude >$$

For example, router interfaces that have *LocationID* 73 are in the country "US", the region "IL", the city "Chicago", with latitude 41.874001273687, and longitude -87.7206104891848.

3.2.1 Choosing Vantage Points, Landmarks and Targets

Because our methodology involves active probing, we consider only a portion of the total set of router interfaces from the October 2014 DRoP dataset. Limiting the set of interfaces makes the experiment more manageable, while also preventing overload of the Ark infrastructure.

Of the 6,041,769 router interfaces that DRoP geolocated to 8,141 locations, we selected the first 20 router interfaces from each location as candidates for geolocation. As some locations do not have router interfaces, and some has less than 20 router interfaces, we obtain only 58,191 candidate router interfaces. Of these 58,191 candidate router interfaces, we tested their suitability for our experiment (Section 3.2.2). We found 26,945 suitable candidate router interfaces. However, 27K candidates will require a total of 162K traceroutes from six vantage points. We therefore further reduced the number of candidates to minimize the load on Ark, and the time required to test the performance of our methodology.

With our methodology outline in Section 3.2.2, we reduce the number of candidate router interfaces to 9,255. We assign 4,638 router interfaces to the landmarks list, and 4,617 router interfaces to the targets list. The procedure is as follows. For each of the 27K candidate router interfaces, we group them according to their locations. From each group, we select at most one candidate router to add to the landmarks list, and at most one candidate router to add to the targets list. The two lists serve as the endpoints for traceroutes (Section 3.3.)

3.2.2 Determining Candidate Router Interface Suitability

Candidate router interfaces are tested for suitability before performing traceroutes from each vantage point. Our methodology performs traceroutes to candidate router interfaces, to record the intermediate hops, and the RTTs for each hop and the destination. Therefore, traceroutes that include one or more unresponsive hops or an unresponsive destination are deemed unsuitable for geolocation. Our suitability requirements reduce 58,191 candidate router interfaces to just 9,255 router interfaces. The significant reduction of candidate router interfaces imposes limitation to our methodology, which we discuss in Section 3.5.

3.3 Performing Trace Routes to Landmarks and Targets

Traceroutes from vantage points to landmarks and to targets are performed to obtain hop and RTT information in order to calculate error distance vectors (Section 3.4.) For each vantage point, we submit two traceroute requests to Ark: one traceroute request from the vantage point to the landmark, and another from the vantage point to the target (Figure 3.1). To generate traceroute information for our geolocation experiment involving six vantage points, we perform $6 \times 4,638 = 27,828$ traceroutes for landmarks, and $6 \times 4,617 = 27,702$ traceroutes for targets. In total, we perform 55,530 traceroutes for our geolocation experiment. The 55K traceroutes excludes the traceroutes that are performed during router interface suitability test.

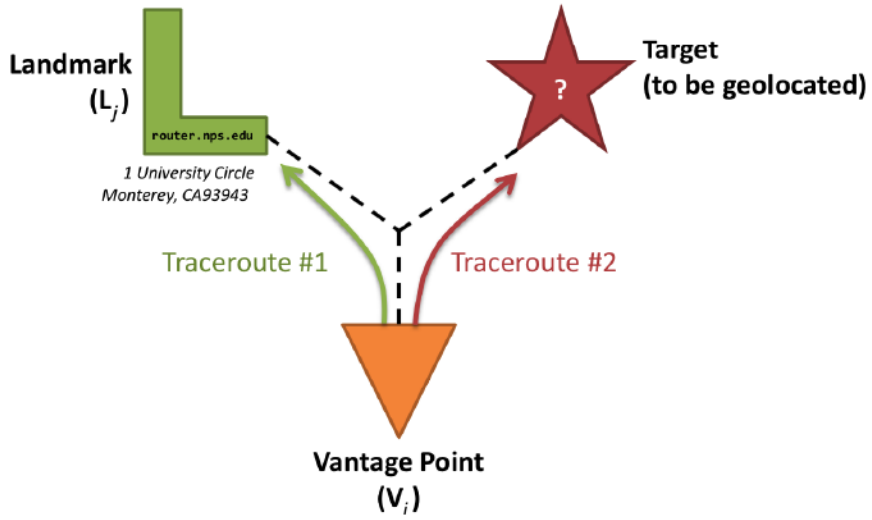

Figure 3.1: For each vantage point, V_i, one traceroute request is submitted from V_i to a landmark, L_j, and another from V_i to the target.

3.4 Generating Estimated Delay Vectors

The estimated delay vector, E, records the estimated delay, D, between a landmark, L, and the target, T, measured from a given vantage point, V. The estimated delay vector of T, E_T, is calculated from the results of two traceroutes: a traceroute from V to L and a traceroute from V to T (Equation 3.1).

$$E_T = < V, L, D > \tag{3.1}$$

D approximates the geographic distance between L and T. D is the sum of the delay, d_1, from the *fork point*, F, to the landmark, and the delay, d_2, from the fork point to the target (Equation 3.2). D thus upper-bounds the possible delay between the landmark in question and the target.

The procedure to determine the fork point, F, is explained in Section 3.4.1.

$$D_{V_i,L_j,T} \approx d_1 + d_2 \tag{3.2}$$

To calculate d_1 we take the difference in the RTT from V to L, and the RTT from V to F (Equation 3.3).

$$d_1 = RTT_{V,L} - RTT_{V,F} \tag{3.3}$$

To calculate d_2 we take the difference in the RTT from V to T, and the RTT from V to F (Equation 3.4).

$$d_2 = RTT_{V,T} - RTT_{V,F} \tag{3.4}$$

Each target, T_i, has a set of estimated delay vectors, S_{T_i} (Equation 3.5). The set of estimated delay vectors is used to determine the nearest landmark, which provides and approximation of the target location (Section 3.4.2).

$$S_{T_i} = \bigcup_{j \in V} \bigcup_{k \in L} < V_j, L_k, D_{V_j,L_k} > \tag{3.5}$$

3.4.1 Determining the Fork Point

The fork point is defined as the *last* hop *before* the two traceroutes from the same vantage point splits into separate paths (Figure 3.2). Algorithm 1 summarizes the procedure to determine the fork point of two traces. For example, if the hops of the first traceroute are $< A, B, C, D, E, F, G, H, Landmark >$, and the hops of the second traceroute are

$< A, B, X, D, E, F, Y, Z, Target >$, then the fork point is the sixth hop.

Algorithm 1 To determine the index of the last hop i.e., fork point before t_1 and t_2 splits into two traces.

```
Let t[i] be i-th hop of trace t.
procedure forkpoint(t1, t2)
    t_short ← SHORTERTRACE(t1, t2)
    t_long ← LONGERTRACE(t1, t2)
    previous ← 0
    fork ← 0
    for i = 0...LENGTH(t_short) do
        if t_short[i] = t_long[i] then
            previous ← i
        else
            if previous + 1 = i then
                fork ← i
            end if
        end if
    end for
    return fork
end procedure
```

By determining the last hop before the two traceroute splits, we mitigate single-hop changes in the two traceroutes, e.g., hop X in the preceding example, which would have caused our methodology to determine the fork point prematurely as the third hop. Single-hop changes in a traceroute may be caused by network load balancers as network load is distributed across multiple router interfaces. As a result of premature fork point, a nearer landmark may be calculated as having a longer estimated delay than a further landmark. This leads to our methodology selecting the further landmark as the nearest landmark, which lowers geolocation accuracy.

3.4.2 Geolocating the Target with Estimated Delay Vectors

With the set of estimated delay vectors for a target, S_T, we determine the *minimum estimated delay vector*, E_{min}, as the estimated delay vector with the minimum estimated delay in the S_T. The condition is defined at Equation 3.6.

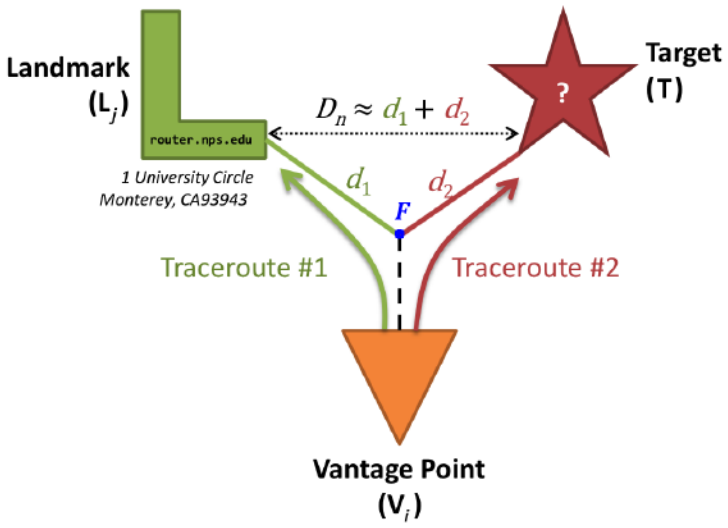

Figure 3.2: F is the fork point of the two traceroutes from vantage point, V_i. The values of d_1 and d_2 are calculated from the F to the landmark, L_j and the target, T, respectively.

$$\forall i \in S_T : D_{E_{min}} \leq D_i \tag{3.6}$$

The location of the target router, T, is approximated by the location of the landmark nearest to T, i.e., the landmark corresponding to the minimum estimated delay vector, E_{min}. For example, the nearest landmark to the target, T, is L_4 in Figure 3.3. Therefore, our methodology geolocates T to the location of L_4.

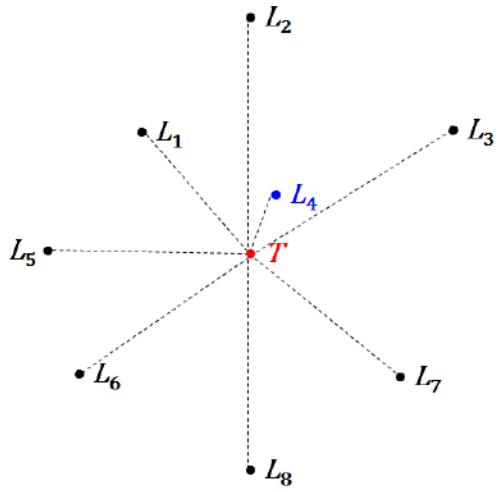

Figure 3.3: Landmark L_4 is the nearest to the target. Therefore, the target geolocation is approximated as the location of L_4.

3.4.3 Negative Estimated Delays

Negative distances may occur due to variations in the RTTs of traceroutes. A traceroute operation initiates multiple Internet Control Message Protocol (ICMP) packets with increasing Time To Live (TTL) values. Due to variations in queuing and processing delays in the network, a further hop such as the destination host, may have a shorter RTT than the hops before it. From Equation 3.3, d_1 will be negative if $RTT_{V,F}$ is larger than $RTT_{V,L}$. Similarly, from Equation 3.4, d_2 will be negative if $RTT_{V,F}$ is larger than $RTT_{V,T}$. From Equation 3.2, negative values for either d_1 or d_2 may cause the estimated delay, D, to be negative.

We illustrate the occurrence of negative distance calculation with two traceroute examples. Table 3.2 shows two traceroutes. The first traceroute is from the vantage point 128.232.97.2 to the landmark 68.86.210.86. The second traceroute is from the same vantage point but to the target 67.178.228.22. The fork point is determined as hop 11. From Equation 3.3, $d_1 = 88.624 - 96.144 = -7.52$. From Equation 3.4, $d_2 = 94.156 - 96.02 = -1.864$. Therefore, from Equation 3.2, the estimated delay, $D = -7.52 - 1.864 = -9.384$.

Estimated delay vectors should not have negative estimated delays and should be discarded. Negative estimated delays are the effects of errors in multiple delay measurements. In

	To Landmark		To Target	
No.	Hops	RTT	Hops	RTT
1	128.232.97.2	4.743	128.232.97.2	58.288
2	193.60.89.5	0.402	193.60.89.5	0.414
3	192.84.5.137	5.957	192.84.5.137	0.428
4	192.84.5.94	0.553	192.84.5.94	0.465
5	146.97.130.1	0.298	146.97.130.1	0.301
6	146.97.37.185	2.638	146.97.37.185	2.641
7	146.97.33.30	19.145	146.97.33.30	5.847
8	146.97.33.10	5.896	146.97.33.10	5.914
9	80.231.60.49	5.896	80.231.60.49	5.906
10	80.231.130.129	79.902	80.231.130.129	82.441
***11**	80.231.130.34	**96.144**	80.231.130.34	**96.02**
12	**216.6.57.2**	95.528	**66.198.70.13**	98.129
13	66.198.111.146	94.138	66.198.70.2	97.377
14	66.198.111.126	91.086	66.198.111.126	97.539
15	66.110.96.5	87.833	66.110.96.5	97.418
16	66.110.96.150	83.801	66.110.96.146	94.795
17	68.86.83.93	82.928	68.86.83.93	90.678
18	68.86.90.22	82.103	68.86.95.166	109.31
19	68.85.62.242	85.04	68.85.63.145	88.894
20	68.86.158.18	86.015	68.86.158.126	87.341
21	68.86.210.86	**88.624**	68.86.158.2	90.227
22	-	-	98.205.28.53	96.678
23	-	-	67.178.228.22	**94.156**

Table 3.2: Traceroute from vantage point 128.232.97.2 to landmark 68.86.210.86, and target 67.178.228.22. At hop 11, the two traceroutes fork.

reality, a packet cannot reach the destination host earlier than any of the intermediate hops. The presence of negative estimated delays may cause our methodology to select the less optimal landmark and produce inaccurate geolocation results.

Our methodology has two ways to handle negative estimated delays. First, we repeated our traceroutes once and considered only the *minimum RTT* for each hop of each traceroutes as the *effective RTT* for calculating estimated delay. This reduces the spikes in the RTT of each hop because our method considers the hop with shortest RTT. Second, we discard estimated delay vectors with negative d_1 or d_2. While calculating the estimated delay for each estimated delay vectors, if d_1 or d_2 is negative, then we assigned it with an arbitrarily

large value of 999,999 ms. This ensures that the resultant estimated delay vector will not be the estimated delay vector with the shortest estimated delay, which has the same effect as discarding the estimated delay vector.

3.4.4 Comparing Geolocation Result with Ground Truth

The haversine formula (Equation 3.7) is used to calculate the distance between two points on the spherical Earth.

$$haversine(\theta) = \sin^2(\frac{\theta}{2}) \tag{3.7}$$

The distance, d, between two points (ϕ_1, λ_1) and (ϕ_2, λ_2) on Earth is calculated with Equation 3.8.

$$d = 2r\arcsin(\sqrt{haversine(\phi_2 - \phi_1) + \cos(\phi_1)\cos(\phi_2)haversine(\lambda_2 - \lambda_1)}) \tag{3.8}$$

$$= 2r\arcsin(\sqrt{\sin^2(\frac{\phi_2 - \phi_1}{2} + \cos(\phi_1)\cos(\phi_2)\sin^2(\frac{\lambda_2 - \lambda_1}{2}))}) \tag{3.9}$$

Where

- d is the distance between two points on the Earth.
- r is the radius of the Earth, approximated to 6,367 km.
- ϕ_1 is the latitude of the first point.
- λ_1 is the longitude of the first point.
- ϕ_2 is the latitude of the second point.
- λ_2 is the longitude of the second point.

From Equation 3.9, we calculate the *error distance*, d, as the geographic distance between the estimated geographic coordinates and its actual geographic coordinates. This metric is used to determine the accuracy of our methodology in Chapter 4. The actual geographic coordinates of the landmarks and targets are obtained from DRoP locations, which we assume as ground truth. As the DRoP locations may be inaccurate, we investigate their

accuracy in Chapter 4.6.

3.5 Limitations

The intermediate hops and destination of each traceroute must be responsive and the traceroute must be complete. Every hop along the path to the destination must reply to the probe and the ICMP response must be returned so that traceroute can determine the forward path interface addresses and RTTs. Each traceroute must be able to reach its destination to be considered complete. The two requirements reduce the number of suitable landmarks and targets for the experiment because intermediate routers and destinations may not respond to probes or return ICMP responses for security reasons. Network administrators may not want to reveal information about their network topology through ICMP probe replies. In Section 3.2.2, our suitability test shows that only 15.9% of the candidate router interfaces are suitable for our methodology.

Because our methodology relies on finding the nearest landmark, our accuracy is tied to the accuracy of the location where we believe the landmark to be. While we have the locations of landmarks from the DRoP dataset, we are uncertain of their accuracy. Huffaker *et al.* performed limited validation of DRoP and found that only 14% of the router interfaces belonging to belwue.de are geolocated outside a 10 km radius of the true location [8]. As our landmarks are obtained from the DRoP results, it is necessary to verify the accuracy of these landmarks. We elaborate on our method to verify ground truth next in Section 3.6.

3.6 Verifying Router Interface Ground Truth with CBG

We used the CBG technique [25] to verify the accuracy of the results of DRoP. CBG relies primarily on end-to-end delay measurements between each vantage point and the target. From Equation 2.1, we see that the end-to-end delay is the sum of four delay components: transmission delay, queuing delay, processing delay, and propagation delay. Of the four delay components, the propagation delay dominates the other components over large geographic distances. We assume that each probe travels over optical fiber between a vantage point and the target so each probe travels at two-thirds the speed of light [25]. As a result, the end-to-end delay between a vantage point and the target limits the furthest possible distance that the probe can possibly reach while propagating at two-thirds the speed of light. While CBG can be inaccurate because mapping latencies to distance may be unreliable,

e.g., due to circuitous paths, CBG provides a strong upper bound distance as it is governed by the physics of the speed of light. Since we use DRoP results as ground truth, we expect DRoP results to be consistent with our CBG results.

CBG calculates the feasible region of a target using RTT measurements from multiple vantage points to the target. We sent delay measurement probes from 22 vantage points from the Ark infrastructure to measure the RTTs between each vantage point and the target. The measured delays are processed by scripts by Tony Tran [33], which determine the possible region of the target. For simplicity, we use the *base line* model for each of our vantage points, V. Assuming the propagation delay of optical fiber is two-thirds the speed of light, it has 1 ms RTT per 100 km. This implies that the base line model, y_V, has a slope, $m_{base} = 0.01$, and the y-intercept of 0, which implies no additive delay (Equation 3.10).

$$y_V = 0.01x + 0 \tag{3.10}$$

3.6.1 Determining the Accuracy of Geographic Coordinates for Each Geographic Location

Incorrect geographic coordinates may be assigned in the DRoP locations. The October 2014 DRoP results consist of 8,141 locations [32]. Huffaker *et al.* assigned geographic coordinates to their dictionary of location hints with the geographic coordinates obtained from the public database, GeoNames [10]. Errors may be introduced when entries in GeoName and DRoP's location hint dictionary are mismatched. Refer to Figure 3.4 for an example of a location with incorrect geographic coordinates, and a location with correct geographic coordinates. Location A is defined with geographic coordinates as shown with a yellow, dotted circle. However, router interfaces R_1 to R_5 assigned to location A are the five yellow dots inside the yellow circle. On the other hand, location B is defined with the geographic coordinates as shown with the blue circle. Router interfaces R_6 to R_{10} are the five blue dots that are assigned to B. They lie within the region of B, which is defined by the blue circle. Therefore, no router interfaces are incorrectly assigned to B.

To evaluate the accuracy of each location, we pick one responsive router interface from each location, and use CBG to determine the possible region of that router interface. Out

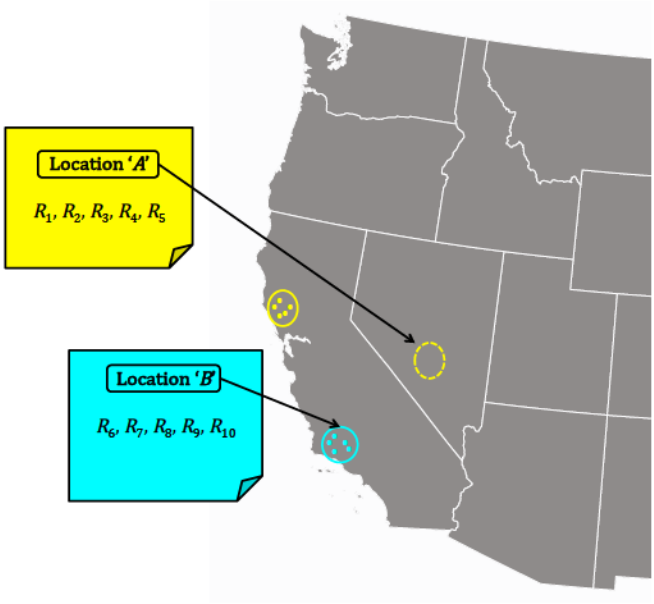

Figure 3.4: Location A has routers R_1 to R_5. A is defined with incorrect geographic coordinates as shown in with the yellow, dotted circle. The router interfaces associated to A are the five dots inside the yellow circle. Location B has routers interfaces R_6 to R_{10}. Router interfaces R_6 to R_{10} are the five blue dots that are assigned to B. They lie within the region of B, which is defined by the blue circle. Therefore, no router interfaces are incorrectly assigned to B.

of the 8,141 global locations that DRoP has determined, we obtain 4,638 locations with at least one responsive router interface. We select 22 available Ark monitors from various parts of the world to carry out the delay measurements. The list of selected vantage points are shown in Table 3.3. We then measured the RTTs for these 4,638 router interfaces from each vantage point, resulting in a total of $4,638 \times 22 = 102,036$ RTTs.

We process the RTTs with CBG to calculate the possible region of the target router interface. The calculated possible region allows us to determine whether DRoP's geolocation of the location of the target router interface lies within the possible region, and calculate the error distance between the target geolocation and the centroid of the possible region.

Vantage Point	Location
ams	Amsterdam, Netherlands
ams2	Amsterdam, Netherlands
bcn	Barcelona, Spain
bwi	Aberdeen, MD, U.S.
cbg	Cambridge, United Kingdom
cgk	Jakarta, Indonesia
cph	Ballerup, Denmark
dac	Darka, Bangladesh
dub	Dublin, Ireland
hel	Espoo, Finland
iad	Chantilly, Virginia, U.S.
jfk	New York, NY, U.S.
lax	Los Angeles, CA, U.S.
mnl	Quezon, The Philippines
per	Perth, Australia
san	San Diego, CA, U.S.
sin	Singapore
sql	Redwood City, CA, U.S.
sjc2	San Jose, CA, U.S.
syd	Sydney, Australia
tpe	Taipei, Republic of China
zrh2	Zurich, Switzerland

Table 3.3: A list of 22 vantage points that are selected for delay measurement in Section 3.6.1.

3.6.2 Determining the Accuracy of Router Interface Assigned to Specific Locations

Ambiguity in router interface DNS PTRs may cause them to be assigned to the wrong location. The October 2014 DRoP results geolocated 6M router interfaces [31] to 8,141 locations. Router interfaces are assigned to locations by matching their hostnames with the geohint database [8]. A router interface may be assigned to the wrong location when its DNS PTR contains substrings that applies to multiple locations. For example, a router interface with the DNS PTR `cr1.chi2ca.sbcglobal.net` may be incorrectly assigned to the location "Chicago, IL" when it should be "Chico, CA". Refer to Figure 3.5 for an example of a location with incorrect router interface assignment, and a location with correct router interface assignment. Location *A* is defined with geographic coordinates as shown in

with the yellow circle. Yet, A has two clusters of router interfaces. The first cluster consists of R_1 to R_5 are shown inside the yellow, dotted circle. The second cluster consists of R_6 to R_{10} are shown inside the yellow circle. The first cluster is likely to have been incorrectly assigned to A. Location B has routers R_6 to R_{10} that cluster in the same region. B is unlikely have incorrectly assigned router interfaces.

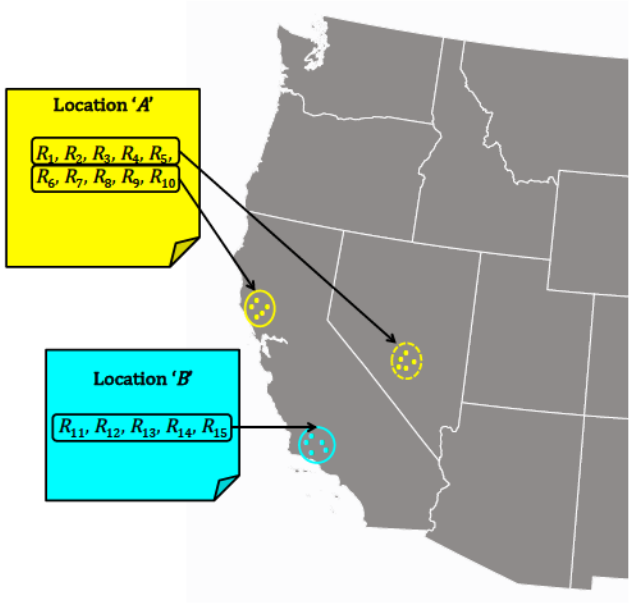

Figure 3.5: Location A is defined with geographic coordinates as shown in with the yellow circle. A has two clusters of router interfaces. The first cluster consists of R_1 to R_5 are shown inside the yellow, dotted circle. The second cluster consists of R_6 to R_{10} are shown inside the yellow circle. The first cluster is likely to have been incorrectly assigned to A. Location B has routers R_6 to R_{10} that cluster in the same region. B is unlikely have incorrectly assigned router interfaces.

To evaluate the accuracy of the assignment of router interfaces to a given location, we sampled at least 20 responsive router interfaces from two locations: "Chicago, IL" and "New York City, NY". We selected two US cities because the largest concentration of Ark vantage points is in North America, and CBG produces the best results when the vantage points are near the target [27]. We select 15 available Ark monitors from North America to carry out the delay measurements. The list of selected vantage points are shown in Table 3.4.

The results of CBG on the selected router interfaces in Chicago and New York City are

Vantage Point	Location
amw	Ames, IA, U.S.
bjc	Broomfield, CO, U.S.
bwi	Aberdeen, MD, U.S.
fnl	Fort Collins, CO, U.S.
iad	Chantilly, Virginia, U.S.
jfk	New York, NY, U.S.
lax	Los Angeles, CA, U.S.
ord	Chicago, IL, U.S.
rno	Reno, NV, U.S.
san	San Diego, CA, U.S.
sea	Seattle, WA, U.S.
sql	Redwood City, CA, U.S.
sjc2	San Jose, CA, U.S.
wbu	Boulder, CO, U.S.
yyz	Toronto, Ontario, Canada

Table 3.4: A list of 15 vantage points in North America that are selected for delay measurement in Section 3.6.2.

possible regions of the router interfaces. We evaluate the possible regions by analyzing the pairwise distances between each centroid of each possible region of router interfaces assigned to Chicago and New York City (Section 3.6.2).

Pairwise Distances

Pairwise distances between each router interface allow us to estimate the proximity between router interfaces. The distribution of pairwise distances may suggest that the router interface belong to geographically-distant clusters. The presence of clusters may identify router interfaces that are incorrectly assigned to a given location.

We approximate the target location with the centroid of the possible region. CBG calculates a possible region of a target with the RTT measurements between the target and each vantage point. We calculated the centroid of this region and assume it as the location of the target.

Finally, we calculate the pairwise distance between each target geolocation. For example, there are routers A, B, and C in location X. We use CBG to geolocate A, B, and C to get locations L_A, L_B, and L_C. Therefore, the pairwise distances for X are between L_A and L_B,

L_A and L_C, and L_B and L_C.

The distribution of the pairwise distances between each geolocation allows us to determine if the router interfaces in a given location belong in the same location. For example, if A, B, and C are in location X, then their pairwise distances should be small. However, if the pairwise distance between A and B is 10 km, while the pairwise distance between A and C, and B and C is 100 km, then either A and B, or C is incorrectly assigned to location X. Table 3.5 illustrates this example.

Router Interface, R_i	Router Interface, R_j	Pairwise Distance between R_i and R_j
A	B	10 km
A	C	100 km
B	C	100 km

Table 3.5: Assume router interfaces A, B, and C are in location X. Then A and B are in one cluster, and C is in another cluster. The separation between the two clusters suggest either A and B, or C are incorrectly assigned to location X.

CHAPTER 4:
Results and Analysis

The objective of this thesis is to describe and evaluate our methodology of geolocating router interfaces with DRoP results. We built the landmark list and target list with suitable router interfaces from DRoP's October 2014 results. We then used the Ark infrastructure to perform traceroutes from six vantage points (Table 3.1) to the landmarks and targets. We calculated the estimated distance vectors for each target, and found the smallest estimated distance, which thereby determined the landmark nearest to the target. The location of each target was predicted to be the location of its nearest landmark. To evaluate the accuracy of the inferences made by our methodology, we calculated the error distances between each geolocated target and its location as inferred by DRoP. Before providing detailed results, we provide an overview of our findings.

Using DRoP as assumed ground truth, the distribution of error distances for the geolocation of 4,617 router interfaces indicates that our methodology is relatively inaccurate (Section 4.1). Half of the geolocated targets have error distances more than $\sim 2,000$ km. A more in-depth analysis however revealed that the inferred geolocations had short estimated RTTs to the nearest landmark, implying that our methodology had selected a nearby landmark. Long error distances and short estimated RTTs suggest that our ground truth, the DRoP geolocations in our case, may be incorrect. As a result, we selected three geolocation results to examine in detail (Section 4.5). We verified that, in these cases, the ground truth was indeed wrong and affected our results.

Among continents, we found that our methodology geolocates North American router interfaces more accurately than Oceanian router interfaces (Section 4.2). Our results show that 72.5% of the North American targets have error distances less than 5,000 km while only 24.8% of the Oceanian targets have error distances less than 5,000 km.

We determined that our methodology is relatively insensitive to the selection of a single vantage point (Section 4.3). However, we found that more vantage points allow our methodology to select nearer landmarks (Section 4.4). For example, when six vantage

points were used, 97.8% of our geolocation had estimated RTTs less than 50 ms. When only the Cambridge, UK vantage point is used, 90.1% of the inferences had estimated RTTs less than 50 ms.

Our analysis suggested that our ground truth is inaccurate. Therefore, we evaluated DRoP's geolocation predictions. We measured RTTs from vantage points to one router interface of each DRoP location, then applied CBG to estimate the location of that router interface. We also calculated the pairwise distances of the CBG-geolocated router interfaces in Chicago, IL, and New York City, NY. Using the subset of DRoP results inclusive of a router in each of DRoP's possible locations, we assessed DRoP's accuracy as compared to CBG. We found that the majority of routers in this subset have a CBG geographic center far from the coordinates given by DRoP. Herein, we use the term *DRoP location* to mean the location of a router interface that we selected, which DRoP mapped to a particular location. This is a one-to-one relationship as defined in Section 3.6.1. Our results show that 50% of 4,638 DRoP locations had distances of more than 1,800 km away from their CBG-geolocated location. Long distances between the geographic coordinates of DRoP locations and its CBG-geolocated location suggests that DRoP's results are inaccurate.

As a specific example, through our pairwise distance calculations, we found router interfaces in Chico, CA incorrectly associated to Chicago, IL (Section 4.6.2). We also found that New York City, NY router interfaces in our ground truth are physically closer to one another than router interfaces in Chicago. Our results show that 90% of DRoP's router interfaces inferred to be in New York City have pairwise distances of less than 22 km. As the accuracy of our methodology depends on the proximity of landmarks to targets, we believe that our methodology geolocates more accurately for router interfaces in New York City than in Chicago.

4.1 Geolocating Global Router Interfaces From Six Vantage Points

The objective of this experiment is to obtain an overview of the accuracy of our methodology. We performed traceroutes from six vantage points (Table 3.1) from each continent to 4,638 landmarks and 4,617 targets (Section 3.2.1). A total of 465 targets did not respond to our traceroute probes. As a result, we successfully geolocated 4,152 target router interfaces

from 4,152 DRoP locations.

The Cumulative Distribution Function (CDF) for the error distances, assuming DRoP ground-truth, of geolocating 4,152 target routers is shown in Figure 4.1. T error distance was computed with Equation 3.9 defined in Section 3.4.4. The distribution of the error distances suggests that our methodology did not accurately geolocate the targets. Half of the geolocated targets have error distances less than 2,420 km, and 20% of the targets have error distances above 10,000 km. Considering the fact that the circumference of the Earth is $\sim 40,000$ km, we see that 20% of the targets have error distances more than a quarter of the circumference of the Earth.

The CDF for the estimated RTTs between the target and its nearest landmark is given in Figure 4.2. The distribution of the estimated RTTs suggests that 75% of nearest landmarks are estimated to have RTTs less than 10 ms. For simplicity, we assume d_{prop} of optical fiber is 200,000 km/s, and d_{trans}, d_{proc}, and d_{queue} are negligible relative to the propagation delay (Equation 2.1). Therefore, $D_{end-to-end} = d_{prop}$. Given 10 ms $\div 2 = 5$ ms of one-way delay, we can thus approximate the upper bound of the distance between the target and the nearest landmark as 200,000 km/s $\times$ 0.005 s = 1,000 km. As the error distance is approximated by estimated one-way delay, we expect the percentage of targets having error distances less than 1,000 km derived from Figure 4.1 and Figure 4.2 to be similar. Yet from Figure 4.1, only 38.4% of the targets have error distances less than 1,000 km. Table 4.1 compares examples of discrepancies in the percentage of targets having error distances less than 500 km, 1,000 km (explained above), 2,000 km, and 5,000 km derived from Figure 4.1 and calculated from Figure 4.2. A possible explanation for the discrepancies is that the actual geographic coordinates for landmarks, targets, or both are incorrect.

Error Distance, d	Percentage of targets with error distances less than d (from Figure 4.1)	Time, t, required to cover d	RTT, r, required to cover d	Percentage of targets with estimated RTTs less than r (from Figure 4.2)
500 km	30%	2.5 ms	5 ms	58%
1,000 km	38%	5 ms	10 ms	75%
2,000 km	48%	10 ms	20 ms	88%
5,000 km	59%	25 ms	50 ms	98%

Table 4.1: Comparison of the percentages of targets with error distances of less than d km, calculated from Figure 4.1 and Figure 4.2.

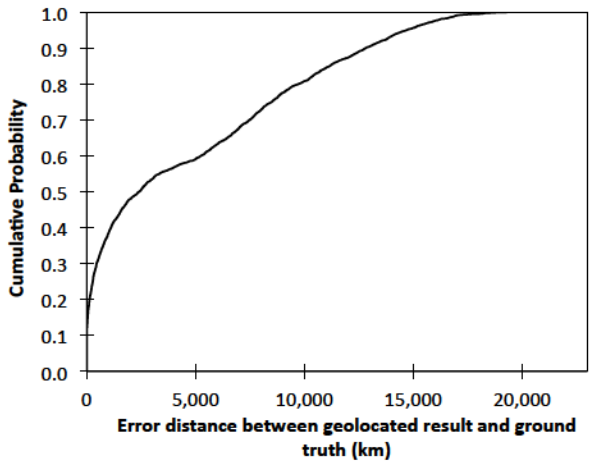

Figure 4.1: CDF for error distances of 4,152 targets from six vantage points.

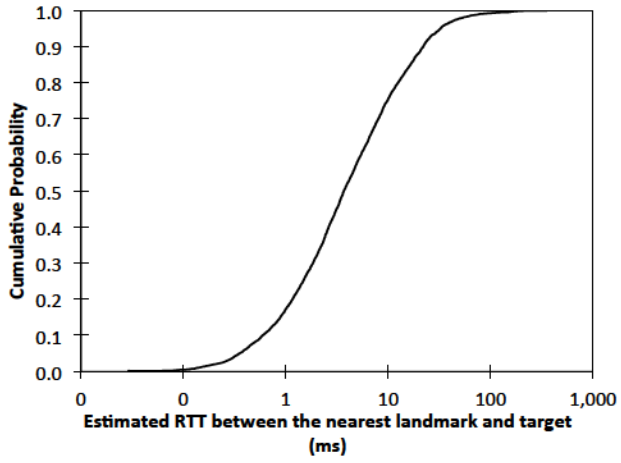

Figure 4.2: CDF for estimated RTTs of 4,152 targets from six vantage points.

4.2 Comparing Geolocation of North American and Oceanian Router Interfaces

The objective of this experiment is to compare the accuracy of geolocating targets in North America and Oceania. This experiment allows us to study the impact of the number of landmarks have on geolocation accuracy as DRoP provided more landmarks in North America than Oceania. We selected 2,032 targets in North America and 137 targets in Oceania for geolocation. We assume the ground truth of the targets is correct, specifically, the selected targets are indeed in the respective country (either North America or Oceania).

The CDF for the error distances of North America and Oceania targets are shown in Figure 4.3. The geolocation results for North American targets resembles the global results (Section 4.1) because the majority of the landmarks and targets involved in the global results are in North America.

The geolocation results for North American targets are more accurate than Oceanian targets: 72.5% of the North American targets have error distances less than 5,000 km while only 24.8% of the Oceanian targets have error distances less than 5,000 km. Figure 4.4 examines the observed latency from the nearest landmark to the targets. We observe that 98.5% of North American and Oceania targets have estimated RTTs of less than 61 ms. While our methodology had found comparably close landmarks for targets in both regions (Figure 4.4), the distribution of error distances for the two regions are drastically different (Figure 4.3). A possible explanation for the difference in accuracy is that the ground truth for Oceanian landmarks, or targets, or both, are inaccurate.

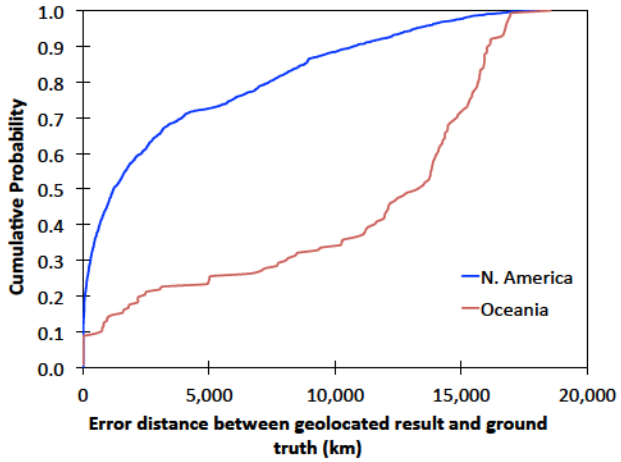

Figure 4.3: CDF for error distances of 2,032 North America targets and 137 Oceania targets.

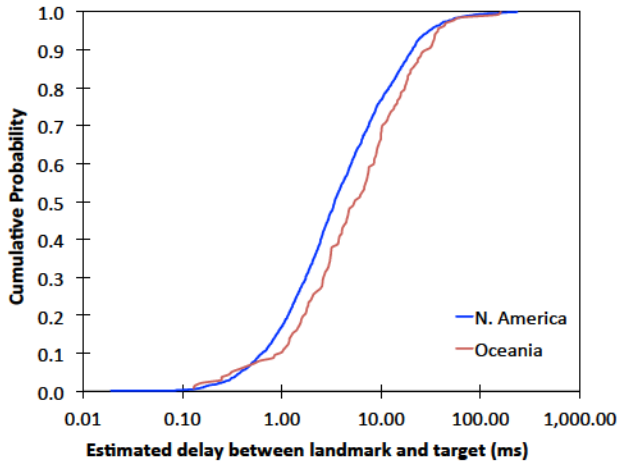

Figure 4.4: CDF for estimated RTTs of 2,032 North America targets and 137 Oceania targets in $\log_{10}$ scale.

4.3 Influence of Vantage Point Location on Geolocation Accuracy

The objective of this experiment is to investigate the impact of the choice of vantage point on our geolocation methodology. In particular, we examine the accuracy of geolocating global targets with traceroutes launched from vantage points in each continent. The Ark monitors that were used as vantage points for this experiment are listed in Table 3.1.

The CDF for the error distances and estimated RTTs of geolocating 4,152 target routers from each of the six vantage points is shown in Figure 4.5 and Figure 4.6 respectively. The CDFs of error distances for each of the six vantage points are similar to one another. This similarity suggests that the choice of vantage points do not improve accuracy. The CDFs of estimated RTTs for each of the six vantage points are also similar to one another. This similarity suggests that the choice of vantage points do not affect the selection of the nearest landmark. Therefore, the accuracy of our methodology may be dependent on other factors such as the number of vantage points involved instead of the location of the vantage points used. A possible reason that geolocation accuracy is insensitive to the choice of vantage point is that the traceroute from a particular vantage point, V_i, may discover a more optimal nearest landmark for a target, T_m, but it also discovers a less optimal nearest landmark for target, T_n, However, the traceroute from vantage point, V_j, discovers a more optimal nearest landmark for T_n.

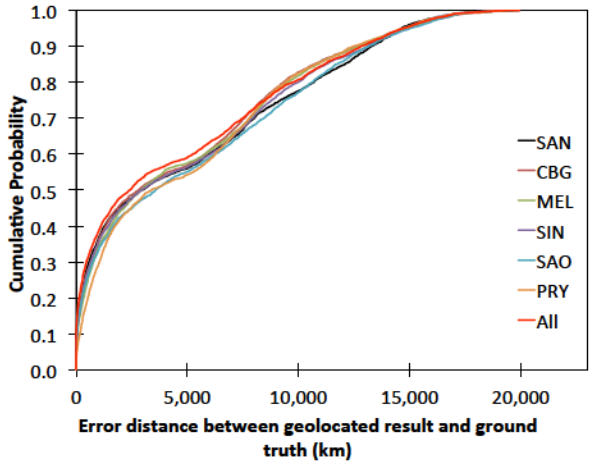

Figure 4.5: CDF for error distances of 4,152 targets from each of the six vantage points, and from all vantage points.

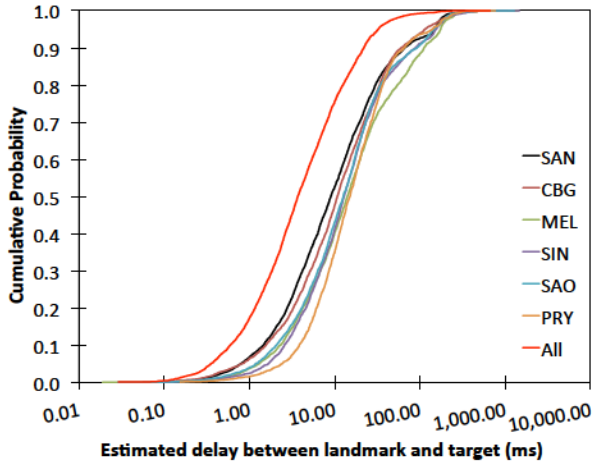

Figure 4.6: CDF for estimated RTTs of 4,152 targets from each of the six vantage points, and from all vantage points, in $\log_{10}$ scale.

4.4 Impact of Single vs. Multiple Vantage Points on Geolocation Accuracy

The objective of this experiment is to investigate the impact of the number of vantage points on the accuracy of our technique. The intuition behind more vantage points improving accuracy is that more vantage points expose our methodology to more traceroute results. By considering more results, the methodology is able to select the most optimal nearest landmark for the given target. This improves the accuracy of geolocation.

Figure 4.5 shows the CDFs for error distances of targets when probes were launched from each of the six vantage points, and when probes were launched from six vantage points. Figure 4.6 shows the CDF for estimated RTTs of targets when probes were launched from each of the six vantage points, and when probes were launched from six vantage points.

The distribution of estimated RTTs shows that performing traceroutes from multiple vantages allows our methodology to pick nearer landmarks. The accuracy improvement can be seen in the CDF in logarithmic scale (Figure 4.6). With traceroute results from six vantage points, our methodology determined 97.8% of landmarks that have estimated RTTs of less than 50 ms from the target. However, with traceroute results from the vantage point in Cambridge, UK, only 90.1% of landmarks have estimated RTTs of less than 50 ms from the target.

The discrepancies between the distributions of error distances and estimated delays may be caused by inaccurate ground truth. Our methodology selects nearer landmarks when more vantage points are used for geolocation, which should lead to lower error distances. Yet, the error distances were not reduced. We attribute this again to inaccurate ground truth of the landmark, target, or both.

4.5 Analyzing Specific Geolocation Results

We selected three specific geolocation results to more deeply investigate and determine the cause of long error distances and estimated RTTs. The first result has a long error distance and estimated RTT. The second result has long error distance but short estimated RTT. The third result has short error distance but high estimated delay.

4.5.1 Long Error Distance and Long Estimated RTT

The first geolocation result that we examine has a long error distance and a long estimated RTT. A long estimated RTTs suggests that our methodology did not find an optimal landmark from the available landmarks to infer the location of the target. A long error distance may be the result of selecting a suboptimal landmark, or that the ground truth for the target, landmark, or both are wrong. We examine the geolocation result for the target with IP address 115.111.183.237 in Table 4.2 where we find an error distance of 14,086 km and an estimated RTT of 283 ms. We queried IP2Location [18], an open IP geolocation database, for an independent estimate of the location of target and nearest landmark.

The ground truth for the nearest landmark and target, which are based on DRoP's result, differ from the result returned by IP2Location as shown in Table 4.3. DRoP determined that the target is in the U.S. while IP2Location geolocated it to India. DRoP determined the nearest landmark is in Papua New Guinea while IP2Location geolocated it to CA, U.S.

If we assume that IP2Location geolocation is correct, then the long estimated RTT of 283 ms is explained by the long distance between the target, which is in India, and the nearest landmark, which is in CA, U.S. The conclusion of this geolocation result is that our methodology failed to find an optimal landmark, resulting in a large error distance.

Result of Geolocation				Location From DRoP	
Error Distance	Estimated RTT	Target Router Interface	Nearest Landmark	Nearest Landmark	Target Router Interface
14,086 km	283 ms	115.111.183.237	137.164.42.242	Port Moresby, Papua New Guinea	Cumberland, RI, U.S.

Table 4.2: The result of geolocating 115.111.183.237.

Target 115.111.183.237 (inpudiidnsprprd01.tata communications.com) Location		Nearest Landmark 137.164.42.242 (dc-pom-csu-lax-dc2-10ge.cenic.net) Location	
By DRoP	By IP2Location	By DRoP	By IP2Location
Cumberland, RI, U.S.	Chetput, Tamil Nadu, India	Port Moresby, Papua New Guinea	Cypress, CA, U.S.

Table 4.3: Comparing the ground truth for the target 115.111.183.237, and the landmark 137.164.42.242 with IP2Location database.

4.5.2 Long Error Distance and Short Estimated RTT

The second geolocation result that we examine has long error distance but short estimated RTT. Geolocation results with long error distances and short estimated RTTs suggest that our methodology had found an optimal landmark but our ground truth includes the wrong location for the landmark, target, or both. We examine the geolocation result for the target with IP address 146.6.137.125, which produced an error distance of 18,276 km and an estimated RTT of 0.379 ms. Table 4.4 summarizes this result. We queried IP2Location for the location of the nearest landmark and target.

The ground truth for the nearest landmark and target, which are based on DRoP's result, are different from the result returned by IP2Location. Table 4.5 compares the differences between DRoP's geolocation and IP2Location's geolocation. DRoP determined the target is in the China, and the nearest landmark is in Chile. IP2Location geolocated the target and the landmark to TX, U.S. DRoP geolocated 146.6.137.125, which has the PTR *ccp*`-test.its.utexas.edu`, to Chile because the PTR contains the substring "ccp", which is the airport code for Concepcion, Chile. Similarly, DRoP geolocated 128.83.10.110, which has the PTR *tnh*`-gi5-5-nocb10.gw.utexas.edu`, to China because the PTR contains the substring "tnh", which is the airport code for in Tonghua Sanyuanpu Airport in Jilin, China.

If we assume that IP2Location geolocation is correct, then the short estimated delay of 0.379 ms is explained by the short distance between the target and its nearest landmark, which are both in Austin, TX, U.S. Furthermore, the target and its nearest landmark have PTRs that are part of the `utexas.edu` domain, which belongs to the University of Texas in Austin, TX. This strongly suggests that the target and its nearest landmark are in Austin,

TX.

The conclusion of this geolocation result is that our methodology found an optimal landmark but the inaccurate ground truth produced a large error distance.

Result of Geolocation				Location From DRoP	
Error Distance	Estimated RTT	Target Router Interface	Nearest Landmark	Nearest Landmark	Target Router Interface
18,276 km	0.379 ms	146.6.137.125	128.83.10.110	Erdaojiang, Jilin, China	Concepcion, Chile

Table 4.4: The result of geolocating 146.6.137.125.

Target 146.6.137.125 (`ccp-test.its.utexas.edu`) Location		Nearest Landmark 128.83.10.110 (`tnh-gi5-5-nocb10.gw.utexas.edu`) Location	
By DRoP	By IP2Location	By DRoP	By IP2Location
Concepcion, Chile	Austin, TX, U.S.	Erdaojiang, Jilin, China	Austin, TX, U.S.

Table 4.5: Comparing the ground truth for the target 146.6.137.125, and the landmark 128.83.10.110 with IP2Location database.

4.5.3 Short Error Distance and Long Estimated RTT

The third geolocation result that we examine has short error distance but long estimated RTT. Geolocation results with short error distances and long estimated RTT suggest that the target's ground truth and its nearest landmark are nearby, but random network delays affected our measurements. We examine the geolocation result for the target with IP address 41.206.162.26, which produced an error distance of 0 km, while the estimated RTT was 110 ms. Table 4.6 summarizes this result. We queried IP2Location for the location of the nearest landmark and target.

The ground truth for the nearest landmark and target, which are based on DRoP's result, are the same as the result returned by IP2Location. Table 4.7 shows that the target and its nearest landmark are geolocated to Tanzania by our methodology and IP2Location.

If we assume that IP2Location geolocation is correct, then our methodology correctly geolocated the target. The estimated RTT should be short because the target and its nearest landmark are in the same city, Dar Es Salam, and they share the same class C IP address. Yet, the estimated RTT is more than 100 ms. We attribute this anomaly to be the result of congestion within the 41.206.162.0/24 network. We believe that repeating traceroutes eliminates sudden increase in delays from d_1 (Equation 3.3) and d_2 (Equation 3.4). As our methodology considers only the minimum error distance, repeating traceroute measurements eliminates this anomaly.

The conclusion of this geolocation result is that our methodology found an optimal landmark and that the long estimated RTT here did not affect the geolocation result.

Result of Geolocation				Location From DRoP	
Error Distance	Estimated RTT	Target Router Interface	Nearest Landmark	Nearest Landmark	Target Router Interface
0 km	110 ms	41.206.162.26	41.206.162.14	Dar Es Salam, Tanzania	Dar Es Salam, Tanzania

Table 4.6: The result of geolocating 41.206.162.26.

Target 41.206.162.26 (`ix-0-0-0-901.core1.41B-Dar-Es-Salam.as6453.net`) Location		Nearest Landmark 41.206.162.14 (`ix-0-0-0-101.core1.41B-Dar-Es-Salam.as6453.net`) Location	
By DRoP	By IP2Location	By DRoP	By IP2Location
Dar Es Salam, Tanzania	Dar Es Salam, Tanzania	Dar Es Salam, Tanzania	Dar Es Salam, Tanzania

Table 4.7: Comparing the ground truth for the target 41.206.162.26, and the landmark 41.206.162.14 with IP2Location database.

4.6 Evaluating Ground Truth with CBG

The first and second geolocation result from Section 4.5 suggest that our ground truth contains inaccuracies. We examine the accuracy of DRoP locations, and the accuracy of router interface association to locations, with the procedures outlined in Section 3.6.1 and Section 3.6.2, respectively. As the result of geolocating a target with CBG is an area, we

approximate the target location with the centroid of the area. For simplicity, we define the centroid of the area as the *result of performing CBG on a given target* in this section.

4.6.1 Accuracy of DRoP Locations

We found that only 54% of 4,638 router interfaces assigned with DRoP locations are also within the CBG possible region. Thus, 46% of the router interfaces are inferred by DRoP to be in regions outside the physical boundaries computed by CBG. This strongly suggests a significant number DRoP locations are incorrect.

The large distances between DRoP's inferred geographic coordinates of a router interface and the result of performing CBG to the same interface suggest that DRoP's results must be inaccurate. The CDF for distances between DRoP's location and the calculated centroid are shown in Figure 4.7.

Our results show that 50% of our ground truth had distances of more than 1,800 km away from their CBG-geolocated location. As CBG results are bounded by physical constraints, the distribution of distances between the each ground truth and their CBG location implies that the majority of our ground truth is far from their true positions.

Our CBG results strongly suggest that the locations defined by DRoP are inaccurate, which explains the root cause of our method's poor geolocation accuracy. When DRoP assigns incorrect geographic coordinates to its locations, these affect the accuracy of our infrastructure geolocation results. Since we used the locations defined by DRoP as ground truth for our landmarks and targets, and the accuracy of our methodology depends on the accuracy of the ground truth. Ground truth locations with incorrect geographic coordinates thus negatively affected our geolocation results.

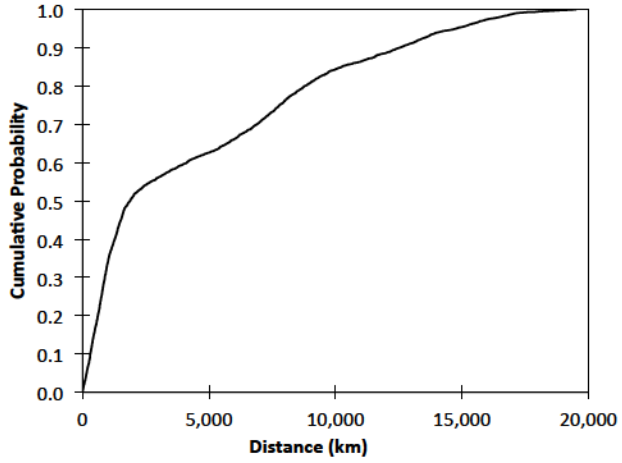

Figure 4.7: CDF for distances between the centroids of CBG geolocations and DRoP's location geographic coordinates.

4.6.2 Accuracy of Router Interface Association to DRoP Locations

Pairwise Distances

With our CBG results, we calculated pairwise distances between 20 router interfaces that DRoP had associated to Chicago, IL, and 40 router interfaces that DRoP had associated to New York City, NY. We selected 40 router interfaces from each location for the suitability test. Of the 40 router interfaces in Chicago, only 20 were suitable. All 40 router interfaces in New York City were suitable. We obtained $20 \times 19 = 380$ pairwise distances for Chicago and $40 \times 49 = 1,560$ pairwise distances for New York City.

We observe that the router interfaces in Chicago are geolocated by CBG to two different locations. From Figure 4.8, 60% of pairwise distances in Chicago are less than 224km, and the remaining 40% pairwise distances are more than 3,000 km. Furthermore, we observe two distinct jumps in cumulative probability for distances between 0 and 250, and between 3,000 and 3,300 km (Figure 4.8). This suggests that router interfaces are grouped into two locations.

We looked up the geographic coordinates of the CBG results for the two groups. The first group with 60% of the router interfaces were geolocated as Milwaukee, WI, and the second group with 40% of the router interfaces be in the Pacific Ocean, 12 miles west of

Santa Barbara, CA. The geolocation results of the second group of router interfaces suggest DRoP had incorrectly associated 40% of our sampled router interfaces to California instead of Wisconsin or Illinois.

IP2Location geolocated the second group of router interfaces to Chico, CA. We conjecture that router interfaces in Chico, CA, may be incorrectly associated to Chicago, IL, due to ambiguous DNS PTR records. For example, one of the router interfaces has the DNS PTR `cr1.chi2ca.sbcglobal.net` and contains the substring "chi", which may have caused DRoP to associate it to Chicago, IL, rather than Chico, CA, which we assume to be the correct location given our corroborating evidence.

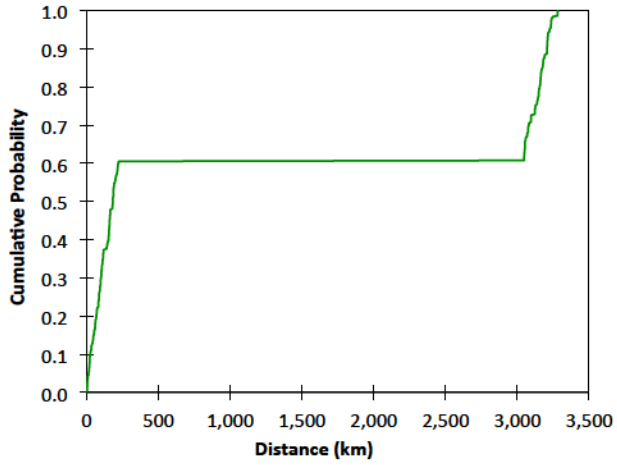

Figure 4.8: CDF for 380 pairwise distances of 20 router interfaces in Chicago, IL.

Our pairwise distance calculations suggest that router interfaces in New York City are close together. From the CDF of pairwise distances of 40 router interfaces in New York City (Figure 4.9), we see that 90% of the sampled router interfaces have pairwise distances of less than 22 km. This implies that the sampled router interfaces are near one another. We looked up IP2Location for five of the 40 router interfaces, which geolocated them to New York City.

From the CDF for pairwise distances of New York City and Chicago (Figure 4.10.), we see that New York City has shorter pairwise distances than Chicago. Assuming that our pairwise distances of the two cities are representative of the inter-router interface distances

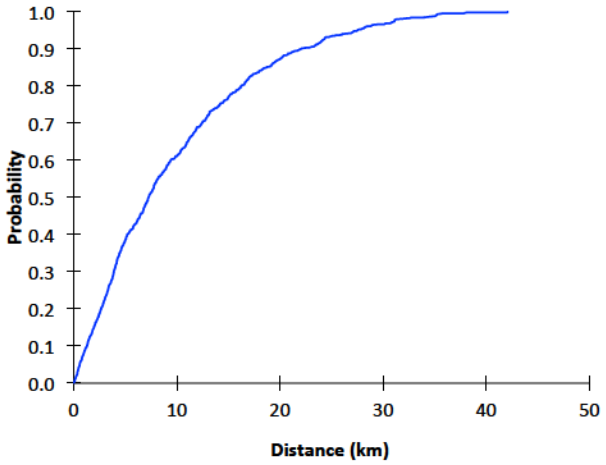

Figure 4.9: CDF for 1,560 pairwise distances of 40 router interfaces in New York City, NY.

of the two cities, then we conjecture that our methodology of IP infrastructure geolocation produces more accurate results for New York City than Chicago. The accuracy of our methodology depends on the proximity of the landmark to the target. If the the landmarks are near one another, then it is likely that our methodology finds nearby landmarks to approximate the location for the target.

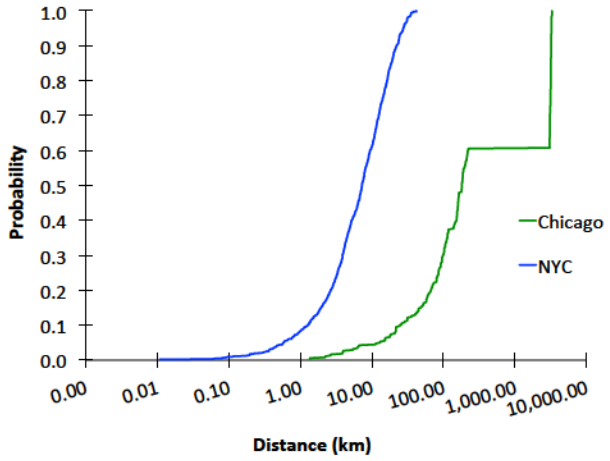

Figure 4.10: CDF for pairwise distances of 20 router interfaces in Chicago, IL, and 40 router interfaces in New York City, NY.

CHAPTER 5:
Conclusion

In this thesis, we proposed our methodology of IP infrastructure geolocation that adapts the street-level, client-independent geolocation technique by Wang *et al.* to geolocate router interfaces that were geolocated by DRoP. We believe that by showing that our methodology accurately geolocates DRoP router interfaces, we could then accurately geolocate router interfaces that DRoP could not.

Our methodology has three main steps. First, we constructed the list of landmarks and targets with the router interface geolocation datasets from DRoP. Next, we used traceroutes from Ark vantage points to actively measure the forward path and forward path round-trip delays to each router interface in the landmark set and each in the target set. Finally, we generated the estimated delay vectors for each target. The predicted geolocation of the target router interface is the location of the landmark from the estimated delay vector with the minimum estimated delay.

The results of our geolocation produced large error distances, which suggested that our methodology is inaccurate. Only half of the geolocated router interfaces had error distances less than 2,420 km while 20% of them had error distances of more than 10,000 km. Yet, we observed that our methodology had selected nearby landmarks according to the inferred landmark to target propagation delay. We therefore suspected that the methodology's inaccurate geolocations may be caused by inaccurate ground truth.

Our results indicated that the location of vantage points do not impact on the accuracy of geolocation. However, using more total vantage points showed an increase in geolocation accuracy.

As our methodology depended on DRoP results as ground truth, we used CBG to evaluate the accuracy of the geographic coordinates assigned to DRoP locations, and the accuracy of router interface association to locations.

We found that 46% of the 4,638 DRoP locations that we had examined are in regions

outside the feasible physical boundaries implied by CBG. This strongly suggests that a significant number DRoP locations are incorrect. Our results also showed that 50% of DRoP locations had distances of more than 1,800 km away from their CBG-geolocated location. As CBG results are bounded by physical constraints, the distribution of distances between each DRoP location and its corresponding CBG location implies the majority of DRoP locations are far from their actual positions. We selected 20 router interfaces in Chicago, IL, and 40 router interfaces in New York City, NY, to be geolocated with CBG. With the results, we performed pairwise distance analysis and possible area intersection. We identified eight router interfaces that were assigned to Chicago should have been assigned to Chico, CA. We observed that 90% of the 40 router interfaces assigned to New York City have pairwise distances of less than 22 km. Our results on the accuracy of DRoP locations, and router interface assignments show that DRoP results are inaccurate.

5.1 Future Work

The results of our IP infrastructure geolocation methodology showed that the accuracy could be improved. To this end, we propose three areas of research to further our work.

Investigate the accuracy of using IP geolocation databases as ground truth. We found that DRoP results were inaccurate, which was problematic given that we relied on DRoP for landmark locations and target ground truth. Potential replacements for DRoP results are IP geolocation databases. It would be interesting to determine the accuracy of our method when using landmark locations and target ground truth as obtained from such geolocation databases.

Explore how accuracy scales with the number of landmarks. Our experiments included 4,638 landmarks to geolocate 4,617 targets. We believe that increasing the number of landmarks would improve accuracy as the probability of finding a landmark closer to the target increases. However, involving more landmarks also increases the number of traceroutes, and the amount of computation required. It would be interesting to observe the geolocation accuracy impact of varying the number of landmarks in this future experiment.

Investigate the feasibility of using traceroute data from CAIDA. Our methodology performed active traceroute measurements to determine router suitability, and to collect per-hop forward round trip latencies. The number of traceroute involved in geolocation

can be significant if the large number of landmarks or vantage points are involved. For example, to geolocate 1 target with 4,000 landmarks from 6 vantage points, a total of $(1+4,000) \times 6 = 24,006$ traceroutes are required. If we double the number of landmarks, and use 82 Ark monitors as vantage points, then a total $(1+8,000) \times 82 = 656,082$ traceroutes are required. As previously collected traceroute datasets are publicly available [34], these costly active measurements might not be necessary. This allows the methodology to skip the router interface suitability test, and the traceroute measurements from each vantage point to each landmark, and to each target. More importantly, the use of large traceroute datasets might contain more traceroutes from different vantage points. As we observed in Section 4.4, more vantage points allow the methodology to select nearer landmarks. We thus conjecture that a larger dataset might lead to more accurate geolocation. It would be interesting to evaluate a variant our methodology that uses the Internet Protocol Version 4 (IPv4) traceroute dataset from CAIDA [34], instead of performing traceroute measurements on Ark. The use of passive traceroute data allows us to increase the scale of our geolocation, which may improve geolocation accuracy.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

[1] A. Chowdhry. (2015). NBC to live-stream Super Bowl XLIX free online without requiring a cable subscription. Forbes. [Online]. Available: http://www.forbes.com/sites/amitchowdhry/2015/01/21/watch-super-bowl-xlix-online/

[2] Y. Shavitt and N. Zilberman, "A structural approach for PoP geo-location," in *INFOCOM IEEE Conference on Computer Communications Workshops, 2010*, March 2010, pp. 1–6.

[3] GeoPriv Working Group. (2014, Mar). GeoPriv status pages. IETF. [Online]. Available: https://tools.ietf.org/wg/geopriv/charters

[4] J. Polk, J. Schnizlein, and M. Linsner, "Dynamic Host Configuration Protocol option for coordinate-based location configuration information," RFC 3825, July 2004. [Online]. Available: http://tools.ietf.org/html/rfc3825

[5] H. Schulzrinne, "Dynamic Host Configuration Protocol (DHCPv4 and DHCPv6) option for civic addresses configuration information," RFC 4776, Nov. 2006. [Online]. Available: http://tools.ietf.org/html/rfc4776

[6] M. Thomson and J. Winterbottom. (2010, Mar.). Discovering the local Location Information Server. [Online]. Available: http://tools.ietf.org/html/draft-ietf-geopriv-lis-discovery-15

[7] M. Zhang, Y. Ruan, V. S. Pai, and J. Rexford, "How DNS misnaming distorts Internet topology mapping." in *ATEC '06 Proceedings of the annual conference on USENIX '06 Annual Technical Conference*, 2006, pp. 34–34.

[8] B. Huffaker, M. Fomenkov *et al.*, "DRoP: DNS-based router positioning," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 3, pp. 5–13, 2014.

[9] L. Daigle. (2004, Sep.). WHOIS protocol specification. RFC 3912. [Online]. Available: http://tools.ietf.org/html/rfc3912

[10] GeoNames. (2015). GeoNames the geographical database. [Online]. Available: http://www.geonames.org

[11] Z. Hu, J. Heidemann, and Y. Pradkin, "Towards geolocation of millions of IP addresses," in *Proceedings of the 2012 ACM conference on Internet measurement conference*. ACM, 2012, pp. 123–130.

[12] D. Moore, R. Periakaruppan, J. Donohoe, and K. Claffy, "Where in the world is netgeo.caida.org?" in *International Networking Conference (INET) '00*. Yokohama, Japan: The Internet Society, Jul 2000.

[13] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 4. ACM, 2001, pp. 173–185.

[14] Y. Shavitt and N. Zilberman, "A study of geolocation databases," *CoRR*, vol. abs/1005.5674, 2010.

[15] HostIP. (2015). Domain IP lookup - Hostip.info. [Online]. Available: http://www.hostip.info/dl/index.html

[16] IPInfoDB. (2015). IPInfoDB free ip address geolocation tools. [Online]. Available: http://www.ipinfodb.com/index.php

[17] MaxMind. (2015). MaxMind - GeoIP2: Industry leading ip intelligence. [Online]. Available: https://www.maxmind.com/en/geoip2-services-and-databases

[18] IP2Location. (2015). IP address geolocation to identify website visitor's geographical location. [Online]. Available: http://ip2location.com

[19] S. Laki, P. Mátray, P. Hága, T. Sebók, I. Csabai, and G. Vattay, "Spotter: A model based active geolocation service," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 3173–3181.

[20] Y. Shavitt and N. Zilberman, "A geolocation databases study," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011.

[21] GeoBytes. (2015). Because everybody's somewhere. [Online]. Available: http://geobytes.com

[22] IPligence. (2015). IP address location of web visitors and geolocation services. [Online]. Available: http://www.ipligence.com

[23] DigitalElement. (2015). NetAcuity Industry-Standard Geolocation. [Online]. Available: http://www.digitalelement.com/solutions/

[24] B. Huffaker, A. Dhamdhere, M. Fomenkov, and K. Claffy, "Toward topology dualism: Improving the accuracy of AS annotations for routers," in *Passive and Active Measurement*, ser. Lecture Notes in Computer Science, A. Krishnamurthy and B. Plattner, Eds. Springer Berlin Heidelberg, 2010, vol. 6032, pp. 101–110. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12334-4_11

[25] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-Based Geolocation of Internet hosts," *Networking, IEEE/ACM Transactions on*, vol. 14, no. 6, pp. 1219–1232, 2006.

[26] B. Eriksson, P. Barford, B. Maggs, and R. Nowak, "Posit: a lightweight approach for IP geolocation," *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 2, pp. 2–11, 2012.

[27] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. ACM, 2006, pp. 71–84.

[28] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards street-level client-independent IP geolocation," in *NSDI*, 2011.

[29] Y. Hyun. (2012, Feb). On-demand IPv4 and IPv6 topology measurements. CAIDA. [Online]. Available: http://www.caida.org/publications/presentations/2012/aims_ark_on_demand/aims_ark_on_demand.pdf

[30] Cooperative Association for Internet Data Analysis (CAIDA). (2015). Archipelago measurement infrastructure. [Online]. Available: http://www.caida.org/projects/ark

[31] B. Huffaker. (2014, Oct). Dataset: ITDK router interface IP addresses. CAIDA. [Online]. Available: http://www.caida.org/~bhuffake/temp/DRoP/ip2locs/ip2locs.zip

[32] B. Huffaker. (2014, Oct). Dataset: ITDK geolocations. CAIDA. [Online]. Available: http://www.caida.org/~bhuffake/temp/DRoP/ip2locs/locations.txt

[33] I. V. Tran, "IPv6 geolocation using latency constraints," master's thesis, Naval Postgraduate School, Monterey, CA, 2014.

[34] Cooperative Association for Internet Data Analysis (CAIDA). (2015). The IPv4 routed /24 topology dataset. [Online]. Available: http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California